A FORMULA FOR PREDICTING THE POPULATION OF THE UNITED STATES.

BY PROF. H. S. PRITCHETT.

Reprinted, by permission, from *Transactions of the Academy of Science, St. Louis*, 1891.

It is often desired to represent by a mathematical equation the law connecting a series of observations for which theory gives no explanation. In such a case ignorance of the physical cause of the phenomena observed does not diminish the accuracy of the computed formula for purposes of prediction, provided the observations are accurate and there are enough of them, and provided the same causes continue to operate.

As the forces giving rise to a series of phenomena become more complicated, the equation which would represent the law connecting the phenomena would generally be correspondingly complicated. When such observed quantities result from a few general causes modified by factors varying among themselves in magnitude and direction, it may be possible to represent the observations fairly well by a comparatively simple equation.

The problem of deriving an equation to represent the law of growth of population in the United States is such a case. The factors entering into this growth, such as birth rate and death rate, immigration and emigration, etc., are more numerous and fluctuating than in older and longer-settled countries. Since, however, the only trustworthy means of predicting the population for the future consists in reasoning from the law of growth in the past, it has seemed to me an interesting question to see how nearly the data already at hand could be represented by a mathematical function.

The data available for this discussion, up to December,

1890, are contained in the ten enumerations of the census from 1790 to 1880 inclusive. The results of these enumerations are given in the following table. The population there given is exclusive of the inhabitants of Alaska and of Indians on reservations.

Year.	Population.	Year.	Population.
1790	3,929,214	1840	17,069,453
1800	5,308,483	1850	23,191,876
1810	7,239,881	1860	31,443,321
1820	9,633,822	1870	38,558,371
1830	12,866,020	1880	50,155,783

A preliminary plat showed that these values could be approximately represented by a parabola, and would be closely represented by an equation of the form: —

$$P = A + B\,t + C\,t^2 + D\,t^3$$

where P represents the population and t the time from some assumed epoch.

Expressing the population in millions and fractions of a million, and the time (t) in decades (census years) counting from 1840, the observations furnish the following 10 equations of condition for determining the constants A, B, C and D: —

$$\begin{array}{l|r}
 & v. \\
A - 5\,B + 25\,C - 125\,D - 3.929 = 0 & +\,0.078 \\
A - 4\,B + 16\,C - 64\,D - 5.308 = 0 & -\,0.038 \\
A - 3\,B + 9\,C - 27\,D - 7.240 = 0 & -\,0.176 \\
A - 2\,B + 4\,C - 8\,D - 9.634 = 0 & -\,0.060 \\
A - B + C - D - 12.866 = 0 & +\,0.119 \\
A - 17.069 = 0 & +\,0.411 \\
A + B + C + D - 23.192 = 0 & +\,0.052 \\
A + 2\,B + 4\,C + 8\,D - 31.443 = 0 & -\,0.982 \\
A + 3\,B + 9\,C + 27\,D - 38.558 = 0 & +\,0.758 \\
A + 4\,B + 16\,C + 64\,D - 50.156 = 0 & -\,0.163
\end{array}$$

Solving by the method of least squares, there result the following normal equations: —

$$
\begin{array}{rrrrrl}
10\,A - & 5\,B + & 85\,C - & 125\,D - & 199.395 & = 0 \\
-\ 5\,A + & 85\,B - & 125\,C + & 1333\,D - & 307.645 & = 0 \\
+\ 85\,A - & 125\,B + & 1333\,C - & 3125\,D - & 1598.197 & = 0 \\
-\ 125\,A + & 1333\,B - & 3125\,C + & 25405\,D - & 3409.531 & = 0
\end{array}
$$

From their solution we obtain the most probable values of A, B, C, and D as follows: —

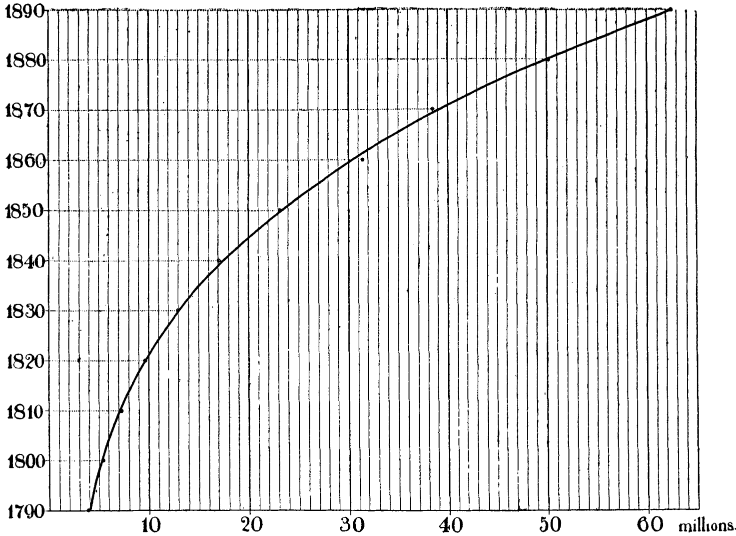
$$
\begin{aligned}
A &= +\ 17.47969 \\
B &= +\ 5.09880 \\
C &= +\ 0.634506 \\
D &= +\ 0.0307275
\end{aligned}
$$

Accordingly, the population "P" for any time "t" would be represented by the equation: —

$$P = 17.47969 + 5.0988\,t + 0.634506\,t^2 + 0.0307275\,t^3 \quad . \quad . \quad .(1)$$

This equation is evidently not what might be called a normal or natural population curve. It has no asymptotes and P becomes zero for a value of t equal to about —9.4, corresponding to the year 1746. For larger negative values of t, P becomes negative. This, however, is what is to be expected from the data used, since the population there given is not the result of a slow natural growth from an original small beginning, but is largely the result of accretions from outside.

How accurately this formula represents the observed values of the population will be seen from the graphical representation of the computed curve which follows. In this plat the axis of Y is the time axis, and the abscissas represent the population expressed in millions. The observed values of the population for each decade are represented by the black dots, and the black-line curve is furnished by formula (1). With the exception of the values for 1860 and 1870, it will be noted that the curve fits the observations with great exactness.



Substituting the values of A, B, C, and D into the equations of condition, there result the residuals given in the column headed "*v*." An examination of these residuals brings out several interesting facts.

The smallness of the residuals, and the consequent close agreement of the formula with the observations, establishes the fact that the general growth of the population has been in the main a regular and orderly one.

There are two residuals which have abnormally large values. These occur in the equations furnished by the Census of 1860 and the Census of 1870. The Census of 1860 shows a population 982,000 greater than the computed value, while the Census of 1870 falls 758,000 short of the computed value. The explanation of these discrepancies is to be found in the effects of the civil war upon the growth of population. The devastating effect of the war would show itself in the Census of 1870 and succeeding years. This effect would be to give a value of the population in 1870 much below that which would be expected. This is precisely what we find to be the

case, the census enumeration in that year falling 758,000 below the computed value. An abnormally small value in 1870 would, of course, have its effect upon the population of succeeding decades, and would give an apparent difference of opposite sign to the observed population in 1860. There is, however, good reason to believe that the value of the population as determined by the census in 1870 is much smaller than the population really was at that time, and there can be little question that the computed value is much nearer the truth than the census determination at that date. The present Superintendent of the Census, Mr. Robert P. Porter, makes the following statement concerning the Census of 1870 (*Census Bulletin No. 12*, Oct. 30, 1890): —

It is well known, the fact having been demonstrated by extensive and thorough investigation, that the Census of 1870 was grossly deficient in the southern states, so much so as not only to give an exaggerated rate of increase of the population between 1870 and 1880 in these states, but to affect very materially the rate of increase in the country at large.

These omissions were not the fault nor were they within the control of the Census Office. The Census of 1870 was taken under a law which the Superintendent, General Francis A. Walker, characterized as "clumsy, antiquated, and barbarous." The Census Office had no power over its enumerators save a barren protest, and this right was even questioned in some quarters. In referring to these omissions the Superintendent of the Tenth Census said in his report in relation to the taking of the census in South Carolina: "It follows as a conclusion of the highest authority either that the Census of 1870 was grossly defective in regard to the whole of the state or some considerable parts thereof, or else that the Census of 1880 was fraudulent." Those, therefore, who believe in the accuracy and honesty of the Tenth Census — and that was thoroughly established — must accept the other alternative offered by General Walker, namely, that the Ninth Census was "grossly defective." What was true of South Carolina was also true, in greater or less degree, of all the southern states.

There is, of course, no means of ascertaining accurately the extent

of these omissions, but in all probability they amounted to not less than 1,500,000. There is but little question that the population of the United States in 1870 was at least 40,000,000, instead of 38,558,-371, as stated.

The computed value just given is 39,316,000; but this is, of course, affected to a certain extent by the error in the Census of 1870, which entered into the computation of formula (1). To compute a value for 1870 which shall be derived from data unaffected by the deficit due to the war, it will be necessary to discuss the observations from 1790 to 1860 alone. The data furnish the following 8 equations of condition: —

	v.
$A - 5B + 25C - 125D - 3.929 = 0$	-0.083
$A - 4B + 16C - 64D - 5.308 = 0$	$+0.166$
$A - 3B + 9C - 27D - 7.240 = 0$	$+0.010$
$A - 2B + 4C - 8D - 9.634 = 0$	-0.090
$A - B + C - D - 12.866 = 0$	-0.136
$A - 17.069 = 0$	$+0.112$
$A + B + C + D - 23.192 = 0$	$+0.083$
$A + 2B + 4C + 8D - 31.443 = 0$	-0.061

Solving by the method of least squares for the value of A, B, C, and D we obtain the following function: —

$$P = 17.1819 + 5.210279\,t + 0.8201904\,t^2 + 0.0623182\,t^3 \quad . \quad . \ (2)$$

How closely this equation fits the observed values will be seen from the table of residuals. These residuals show that during the 70 years from 1790 to 1860 the growth of population followed the law expressed by equation (2) very accurately, and also that this rate of growth was more rapid than that of later decades. Had this rate of growth continued to 1870, the population would have amounted at that time to 41,877,100. The diminution during the decade due to those actually killed, to lessened immigration and decreased birth rate, cannot be stated with exactness, but probably approximates 1,700,000. After deducting this loss it does not seem

possible that the population in 1870 could have been less than 40,000,000, a result entirely in accordance with the conclusions arrived at by the last two Superintendents of the Census.

Had the population continued to grow after 1860 at the same rate as before, we should have had in 1890 a population of over 71 millions, about nine millions more than we really have. It is scarcely possible that the whole of this difference is chargeable to the war, but is probably due in part to a diminishing birth rate.

PROBABLE ERROR.

Assuming the formula correct, there results from the probable error of a single determination of the population ±0.367, expressed as a fraction of a million.

This error contains, of course, both the error of the formula and the error of the census enumeration. Assuming A, B, C, and D as independent quantities, we obtain for their probable errors the following values:—

Probable error of A $= \pm\ 0.179$
Probable error of B $= \pm\ 0.127$
Probable error of C $= \pm\ 0.0178$
Probable error of D $= \pm\ 0.0066$

From these values, expressing P as a function of A, B, C, and D, its probable error may be computed at any time. This probable error would remain a small per cent of the computed population.

VALUE OF THE FORMULA FOR PREDICTION.

How closely formula (1) will continue to represent the growth of population during future decades depends, of course, upon the continuance of the same conditions of growth. A decided change in the birth rate, or rate of immigration, or a destructive war, would bring out a large discrepancy between the computed and observed values. A fair test of the formula is found by computing the population for 1890. According to the formula, we should expect in 1890 a

population of 62,677,280. The Census Bureau has within the last few weeks finished its count of the population in 1890, obtaining the result 62,622,280. The agreement between these two results is all that could be desired, the difference of 55,000 being within the limit of error of both the formula and the census count.

The general law governing the increase of population, as usually stated, is that, when not disturbed by extraneous causes, such as wars, pestilences, immigration, emigration, etc., the increase of population goes on at a constantly diminishing rate. By this it is meant that the percentage of increase from decade to decade diminishes. The law of growth expressed by equation (1) involves such a decrease in the percentage of growth.

Differentiating equation (1) we have

$$\frac{\frac{dP}{dt}}{P} = \frac{B + 2\,C\,t + 3\,D\,t^2}{A + B\,t + C\,t^2 + D\,t^3}$$

which diminishes as t increases, and approaches zero as t approaches infinity. In 1790 the percentage of increase per decade was 32 per cent; in 1880, 24 per cent; in 1990 will be 13 per cent, and in 1000 years will have sunk to a little less than 3 per cent.

In order to include all available data, I have re-solved for A, B, C, and D including the data of 1890. This would yield the following 11 equations of condition :—

	$v.$
$A - 5\,B + 25\,C - 125\,D - 3.9292 = 0$	$+\ 0.083$
$A - 4\,B + 16\,C - 64\,D - 5.3085 = 0$	$-\ 0.041$
$A - 3\,B + 9\,C - 27\,D - 7.2399 = 0$	$-\ 0.181$
$A - 2\,B + 4\,C - 8\,D - 9.6338 = 0$	$-\ 0.065$
$A - B + C - D - 12.8660 = 0$	$+\ 0.119$
$A - 17.0695 = 0$	$+\ 0.415$
$A + B + C + D - 23.1919 = 0$	$+\ 0.058$
$A + 2\,B + 4\,C + 8\,D - 31.4433 = 0$	$-\ 0.975$
$A + 3\,B + 9\,C + 27\,D - 38.5584 = 0$	$+\ 0.754$
$A + 4\,B + 16\,C + 64\,D - 50.1558 = 0$	$-\ 0.181$
$A + 5\,B + 25\,C + 125\,D - 62.6222 = 0$	$+\ 0.012$

These yield the following normal equations: —

$$\begin{aligned}
+\ 11.0\,A + 0.0\,B + 110.0\,C + 0.0\,D - 262.017 &= 0 \\
0.0\,A + 110.0\,B + 0.0\,C + 1958.0\,D - 620.753 &= 0 \\
+\ 110.0\,A + 0.0\,B + 1958.0\,C + 0.0\,D - 3163.765 &= 0 \\
0.0\,A + 1958.0\,B + 0.0\,C + 41030.0\,D - 11237.254 &= 0
\end{aligned}$$

From which result the following values of A, B, C, and D: —

$A = 17.4841$ $B = 5.1019363$ $C = +0.6335606$ $D = +0.0304086$

and the population (P) at any decade (t) after 1840 will be given by the equation,

$$P = 17.4841 + 5.1019363\,t + 0.6335606\,t^2 + 0.0304086\,t^3 \quad . \quad . \quad (3)$$

This formula, being the most probable result deducible from all the data, forms the best basis at hand for predicting the population of the future. In the course of time it is to be expected that this will depart more and more from the observed values, but for the next hundred years will doubtless represent the growth of population within a small percentage of error. Carrying forward the computation, we obtain to the nearest thousand the following values for subsequent dates: —

Year.	Computed Population.	Year.	Computed Population.
1900	77,472,000	1970	257,688,000
1910	94,673,000	1980	296,814,000
1920	114,416,000	1990	339,193,000
1930	136,887,000	2000	385,860,000
1940	162,268,000	2100	1,112,867,000
1950	190,740,000	2500	11,856,302,000
1960	222,067,000	2900	40,852,273,000

It would be interesting to discuss in a similar manner the population of some country like France, in which the growth has been but little affected by emigration. It is the intention of the author to do this as soon as the data are available.

It may be said of the results of the whole discussion that they confirm in a general way, and as far as they go, the accuracy of the Eleventh Census.